Geometric Variance Reduction in Markov Chains: Application to Value Function and Gradient Estimation

Rémi Munos

Journal of Machine Learning Research 7 (2006) 413–427 Submitted 5/05; Revised 1/06; Published 2/06

Geometric Variance Reduction in Markov Chains: Application to Value Function and Gradient Estimation

Rémi Munos REMI.MUNOS@POLYTECHNIQUE.FR
Centre de Mathématiques Appliquées
Ecole Polytechnique
91128 Palaiseau, France

Editor: Shie Mannor

Abstract

We study a variance reduction technique for Monte Carlo estimation of functionals in Markov chains. The method is based on designing *sequential control variates* using successive approximations of the function of interest V. Regular Monte Carlo estimates have a variance of $O(1/N)$, where N is the number of sample trajectories of the Markov chain. Here, we obtain a geometric variance reduction $O(\rho^N)$ (with $\rho < 1$) up to a threshold that depends on the approximation error $V - \mathcal{A}V$, where $\mathcal{A}$ is an *approximation operator* linear in the values. Thus, if V belongs to the right approximation space (i.e. $\mathcal{A}V = V$), the variance decreases geometrically to zero.

An immediate application is value function estimation in Markov chains, which may be used for policy evaluation in a policy iteration algorithm for solving Markov Decision Processes.

Another important domain, for which variance reduction is highly needed, is gradient estimation, that is computing the sensitivity $\partial_\alpha V$ of the performance measure V with respect to some parameter α of the transition probabilities. For example, in policy parametric optimization, computing an estimate of the policy gradient is required to perform a gradient optimization method.

We show that, using two approximations for the *value function* and the *gradient*, a geometric variance reduction is also achieved, up to a threshold that depends on the approximation errors of both of those representations.

1. Introduction

We consider a Markov chain over a finite state space $\mathcal{X}$ defined by the transition matrix P. Note that although we consider a finite state space for simplicity, this work can be easily extended to the case of infinite state spaces (countable or continuous). Write $X(x)$ a trajectory $(x_t)_{t\geq 0}$ starting at a state $x_0 = x$. Let $\Psi(r, X(x))$ be a functional that depends on some function $r : \mathcal{X} \to \mathbb{R}$ and the trajectory $X(x)$, and write $V(x)$ the expectation of the functional that we wish to evaluate:

$$V(x) = \mathbb{E}[\Psi(r, X(x))]. \tag{1}$$

Here, the quantity of interest V is expressed in terms of a **probabilistic representation**, as an expectation of a functional that depends on trajectories. We will consider a functional $\Psi(r,\cdot)$ that is linear in r, and such that its expectation V may equivalently be expressed in terms of a **solution to a linear system**

$$\mathcal{L}V = r, \tag{2}$$

(where r and V are considered as column vectors) with $\mathcal{L}$ an invertible linear operator (matrix).

Such an example of Ψ is the sum of discounted rewards r received along the trajectory:

$$\Psi(r, X(x)) = \sum_{t \geq 0} \gamma^t r(x_t). \tag{3}$$

with $0 < \gamma < 1$ being a discount factor. In that case, V is the solution to the Bellman equation (2) with $\mathcal{L} = I - \gamma P$. Indeed, using matrix notations, $V = \sum_{t \geq 0} \gamma^t P^t r = (I - \gamma P)^{-1} r$.

A regular Monte-Carlo (MC) method would estimate $V(x)$ by sampling N independent trajectories $\{X^n(x)\}_{1 \leq n \leq N}$ starting from x and calculate the average $\frac{1}{N} \sum_{n=1}^{N} \Psi(r, X^n(x))$. The variance of such an estimator is of order $1/N$. Variance reduction is crucial since the numerical approximation error of the quantity of interest is directly related to the variance of its estimate.

Variance reduction techniques include importance sampling, correlated sampling, control variates, antithetic variates and stratified sampling, see e.g. (Hammersley and Handscomb, 1964; Halton, 1970). Geometric variance reduction rates have been obtained by processing these variance reduction methods iteratively, the so-called *sequential* (or *recursive*) *Monte-Carlo*. Examples include adaptive importance sampling (Kollman et al., 1999) and what Halton called the "Third Sequential Method" (Halton, 1994) based on sequential correlated sampling and control variates. This approach has been recently developed in (Maire, 2003) for numerical integration and, more related to our work, applied to (continuous time) Markov processes in (Gobet and Maire, 2005).

The idea is to replace the expectation of $\Psi(r, \cdot)$ by the expectation of $\Psi(r - \mathcal{L}W, \cdot)$ for some function W close to V. From the linearity of Ψ and the equivalence between the representations (1) and (2), for any W, one has

$$V(x) = W(x) + \mathbb{E}[\Psi(r - \mathcal{L}W, X(x))].$$

Thus, if W is a good approximation of V, the residual $r - \mathcal{L}W$ is small, and the variance is low.

In the sequential method described in this paper, we use successive approximations V_n of V to estimate by Monte Carlo a correction E_n using the residual $r - \mathcal{L}V_n$ in Ψ, which is used to process a new approximation V_{n+1}. We consider an approximation operator $\mathcal{A}$ that is *linear in the values*. We show that (for enough sample trajectories at each iteration) the variance of the estimator has a geometric rate ρ^N (with $\rho < 1$, and N the total number of sampled trajectories) until some threshold is reached, whose value is related to the approximation error $\mathcal{A}V - V$.

An interesting extension of this idea concerns the estimation of the gradient $\partial_\alpha V$ of V with respect to (w.r.t.) some parameter α of the transition matrix P. A useful application of such sensitivity analysis appears in policy gradient estimation. An optimal control problem may be approximated by a parametric optimization problem in a given space of parameterized policies. Thus, the transition matrix P depends on some (possible multidimensional) policy parameter α. In order to apply gradient methods to search for a local maximum of the performance in the parameter space, one wishes to estimate the policy gradient, i.e. the sensitivity $Z = \partial_\alpha V$ of the performance measure with respect to α. The gradient may be expressed as an expectation $Z(x) = \mathbb{E}[\Phi(r, X(x))]$, using the so-called *likelihood ratio* or *score method* (Reiman and Weiss, 1986; Glynn, 1987; Williams, 1992; Baxter and Bartlett, 2001; Marbach and Tsitsiklis, 2003). The gradient Z is also the solution to a linear system

$$\mathcal{L}Z = -\partial_\alpha \mathcal{L}\, \mathcal{L}^{-1} r = -\partial_\alpha \mathcal{L} V. \tag{4}$$

(note that the derivative operator ∂_α only applies to $\mathcal{L}$). Indeed, since V solves $V = \mathcal{L}^{-1} r$, we have $Z = \partial_\alpha V = -\mathcal{L}^{-1} \partial_\alpha \mathcal{L}\, \mathcal{L}^{-1} r$. For example, in the infinite horizon, discounted case (3), we have

$\mathcal{L} = I - \gamma P$, thus $\partial_\alpha \mathcal{L} = -\gamma \partial_\alpha P$ and

$$Z = \gamma (I - \gamma P)^{-1} \partial_\alpha P \, (I - \gamma P)^{-1} r = \sum_{t \geq 0} \gamma^{t+1} P^t \partial_\alpha P \sum_{s \geq 0} \gamma^s P^s r.$$

The functional Φ may thus be defined as

$$\Phi(r, X(x)) \quad = \quad \sum_{t \geq 0} \gamma^{t+1} \frac{\partial_\alpha P(x_t, x_{t+1})}{P(x_t, x_{t+1})} \sum_{s \geq 0} \gamma^s r(x_{s+t+1}), \tag{5}$$

which may be rewritten as

$$\Phi(r, X(x)) = \sum_{t \geq 0} \gamma^t r(x_t) \sum_{s=0}^{t-1} \frac{\partial_\alpha P(x_s, x_{s+1})}{P(x_s, x_{s+1})}.$$

We show that, using two approximations V_n and Z_n of the *value function* and the *gradient*, a geometric variance reduction is also achieved, up to a threshold that depends on the approximation errors of both of those representations.

Numerical experiments on a simple Gambler's ruin problem illustrate the approach.

2. Value Function Estimation

We first describe the approximation operator *linear in the values* considered here, then describe the algorithm, and state the main result on geometric variance reduction.

2.1 Approximation Operator $\mathcal{A}$

We consider a fixed set of J *representative states* $\mathcal{X}_J := \{x_j \in \mathcal{X}\}_{1 \leq j \leq J}$ and *basis functions* $\{\phi_j : \mathcal{X} \to \mathbb{R}\}_{1 \leq j \leq J}$. The linear approximation operator $\mathcal{A}$ maps any function $W : \mathcal{X}_J \to \mathbb{R}$ to the function $\mathcal{A}W : \mathcal{X} \to \mathbb{R}$, according to

$$\mathcal{A}W(x) = \sum_{j=1}^{J} W(x_j) \phi_j(x). \tag{6}$$

With a slight abuse of notation, for any function $W : \mathcal{X} \to \mathbb{R}$, we define $\mathcal{A}W : \mathcal{X} \to \mathbb{R}$ similarly from the values of W at $\mathcal{X}_J$. This kind of function approximation includes:

- **Linear approximation**, for example with *Spline*, *Polynomial*, *Radial Basis*, *Fourier* or *Wavelet* decomposition. $\mathcal{A}W$ is the projection of a function W onto the space spanned by a set of functions $\{\psi_k : \mathcal{X} \to \mathbb{R}\}_{1 \leq k \leq K}$, i.e. the function minimizing some norm (induced by a discrete inner product $\langle f, g \rangle := \sum_{j=1}^{J} \mu_j f(x_j) g(x_j)$, for some distribution μ over $\mathcal{X}_J$):

  $$\min_{\alpha \in \mathbb{R}^K} \left|\left| \sum_{k=1}^{K} \alpha_k \psi_k - W \right|\right|^2.$$

  The solution α solves the linear system $A\alpha = b$ with A an $K \times K$-matrix of elements $A_{kl} = \langle \psi_k, \psi_l \rangle$ and b a K-vector of components $b_k = \langle W, \psi_k \rangle$. Thus $\alpha_k = \sum_{l=1}^{K} A_{kl}^{-1} \sum_{j=1}^{J} \mu_j \psi_l(x_j) W(x_j)$ and the best fit $\sum_{k=1}^{K} \alpha_k \psi_k$ is thus of type (6) with

  $$\phi_j(x) = \mu_j \sum_{k=1}^{K} \sum_{l=1}^{K} A_{kl}^{-1} \psi_l(x_j) \psi_k(x). \tag{7}$$

- **Non-parametric approximation**, such as *k-nearest neighbors* (where $\phi_j(x) = \frac{1}{k}$ if x has x_j as one of its $k-$nearest neighbors, and $\phi_j(x) = 0$ otherwise), *locally weighted learning* and *Kernel regression* (Atkeson et al., 1997; Hastie et al., 2001), where functions similar to (7) may be derived (with the matrix A being dependent on x through the kernel), and *Support Vector Regression* (when using a quadratic loss function) (Vapnik et al., 1997; Vapnik, 1998).

2.2 The Algorithm

We assume the equivalence between the probabilistic interpretation (1) and the representation as solution to the linear system (2), i.e. for any function $f : \mathcal{X} \to \mathbb{R}$,

$$f(x) = \mathbb{E}[\Psi(\mathcal{L}f, X(x))]. \tag{8}$$

We consider successive approximations $V_n \in \mathbb{R}^J$ of V defined at the states $\mathcal{X}_J = (x_j)_{1\leq j\leq J}$ recursively:

- We initialize $V_0(x_j) = 0$.
- At stage n, we use the values $V_n(x_j)$ to provide a new estimate of $V(x_j)$. Let $E_n(x_j) := V(x_j) - \mathcal{A}V_n(x_j)$ be the approximation error at the states $(x_j)_{1\leq j\leq J}$. From the equivalence property (8), we have: $\mathcal{A}V_n(x) = \mathbb{E}[\Psi(\mathcal{L}\mathcal{A}V_n, X(x))]$. Thus, by linearity of Ψ w.r.t. its first variable,
$$E_n(x_j) = \mathbb{E}[\Psi(r - \mathcal{L}\mathcal{A}V_n, X(x_j))].$$
Now, we use a Monte Carlo technique to estimate $E_n(x_j)$ at each representative state x_j, using M trajectories $(X^{n,m}(x_j))_{1\leq m\leq M}$: we calculate the average
$$\widehat{E_n}(x_j) := \frac{1}{M}\sum_{m=1}^{M} \Psi(r - \mathcal{L}\mathcal{A}V_n, X^{n,m}(x_j)),$$
and define the new approximation at the states $\mathcal{X}_J$:
$$V_{n+1}(x_j) := \mathcal{A}V_n(x_j) + \widehat{E_n}(x_j). \tag{9}$$

Remark 1 *Notice that there is a slight difference between this algorithm and that of (Gobet and Maire, 2005), which may be written*

$$V_{n+1}(x_j) = V_n(x_j) + \mathcal{A}\Big[\frac{1}{M}\sum_{m=1}^{M} \Psi(r - \mathcal{L}V_n, X^{n,m}(x_j))\Big].$$

Our formulation enables us to avoid the assumption of the idempotent property for $\mathcal{A}$ (i.e. that $\mathcal{A}^2 = \mathcal{A}$) which does not hold in general in non-parametric approximation (e.g. in k-nearest neighbors, for $k \geq 2$) and guarantees that V_n is an unbiased estimate of V, for all n, as showed in the next paragraph.

2.3 Properties of the Estimates V_n

We write the conditional expectations and variances:

$$\mathbb{E}^n[Y] = \mathbb{E}[Y|X^{p,m}(x_j), 0 \leq p < n, 1 \leq m \leq M, 1 \leq j \leq J]$$

and $\mathrm{Var}^n[Y] = \mathbb{E}^n[Y^2] - (\mathbb{E}^n[Y])^2$. We have the following properties on the estimates:

Expectation of V_n. From the definition (9),

$$\mathbb{E}^n[V_{n+1}(x_j)] = \mathcal{A}V_n(x_j) + E_n(x_j) = V(x_j).$$

Thus $\mathbb{E}[V_n(x_j)] = V(x_j)$ for all $n \geq 1$: the approximation $V_n(x_j)$ is an unbiased estimate of $V(x_j)$.

Variance of V_n. Write $v_n = \sup_{1\leq j\leq J} \mathrm{Var}\, V_n(x_j)$. The following result (whose proof is provided in Appendix A) expresses that for large enough values of M, the variance decreases geometrically with n.

Theorem 2 *We have*

$$v_{n+1} \leq \rho_M v_n + \frac{2}{M}\mathcal{V}_\Psi(V - \mathcal{A}V) \tag{10}$$

with $\rho_M = \frac{2}{M}\left(\sum_{j=1}^J \sqrt{\mathcal{V}_\Psi(\phi_j)}\right)^2$, *using the notation*

$$\mathcal{V}_\Psi(f) := \sup_{1\leq j\leq J} \mathit{Var}\, \Psi(\mathcal{L}f, X(x_j)).$$

Thus, for large enough values of M, (i.e. whenever $\rho_M < 1$), $(v_n)_n$ decreases geometrically at rate ρ_M, up to the threshold

$$\limsup_{n\to\infty} v_n \leq \frac{1}{1-\rho_M}\frac{2}{M}\mathcal{V}_\Psi(V - \mathcal{A}V).$$

If V belongs to the space of functions that are representable by $\mathcal{A}$, i.e. $\mathcal{A}V = V$, then the variance geometrically decreases to 0 at rate ρ^N with $\rho := \rho_M^{1/M}$ and N being the total number of sample trajectories per state x_j (i.e. N is the product of the number n of iterations by the number M of trajectories per iteration and state x_j).

Notice that the threshold depends on the variance of Ψ for the function $\mathcal{L}(V - \mathcal{A}V) = r - \mathcal{L}\mathcal{A}V$, the residual of the representation (by $\mathcal{A}$) of V. Notice also that this threshold depends on $V - \mathcal{A}V$ only at states reached by the trajectories $\{X(x_j)\}_{x_j\in\mathcal{X}_J}$: a uniform (over the whole domain) representation of V is not required.

Of course, once the threshold is reached, a further convergence of $O(1/N)$ can be obtained thereafter, using regular Monte Carlo.

2.4 Example: The Infinite Horizon, Discounted Case

Let us illustrate the sequential control variates algorithm to value function estimation in Markov chains in the infinite horizon, discounted case (3). The value function

$$V(x) = \mathbb{E}\Big[\sum_{t\geq 0} \gamma^t r(x_t)\Big]$$

solves Bellman's equation: $V = r + \gamma PV$, which may be written as the linear system (2) with $\mathcal{L} = I - \gamma P$.

In the previous algorithm, at stage n, the approximation error $E_n(x_j) = V(x_j) - \mathcal{A}V_n(x_j)$ is therefore the expectation

$$E_n(x_j) = \mathbb{E}\Big[\sum_{t\geq 0} \gamma^t [r(x_t) - \mathcal{A}V_n(x_t) + \gamma P\mathcal{A}V_n(x_t)] | x_0 = x_j\Big]. \tag{11}$$

We notice that the term $r - \mathcal{A}V_n + \gamma P\mathcal{A}V_n$ is the *Bellman residual* of the approximation $\mathcal{A}V_n$. The estimate thus has zero variance if this approximation happens to be the value function. Following the algorithm, the next approximation V_{n+1} is defined by (9) with $\widehat{E_n}(x_j)$ being a Monte Carlo estimate of (11).

Remark 3 *Note that the expectation operator P may not be easy to process. In model-free learning, it would be interesting to replace the term $P\mathcal{A}V_n(x_t)$ by $\mathcal{A}V_n(x_{t+1})$ in (11) leaving the expectation unchanged. However, this would introduce some additional variance that annihilates the benefit of the technique.*

Nevertheless, the term $P\mathcal{A}V_n$ may actually be computed as $\mathcal{A}'V_n$, where $\mathcal{A}'$ is an approximation operator defined by another set of basis functions $\{\phi'_j := P\phi_j\}_{1\leq j\leq J}$ (i.e. $\phi'_j(x) := \sum_{y\in\mathcal{X}} P(x,y)\phi_j(y)$, $1 \leq j \leq J$). Indeed, for any $W : \mathcal{X}_J \to \mathbb{R}$,

$$P\mathcal{A}W(x) = \sum_{y\in\mathcal{X}} P(x,y) \sum_{j=1}^{J} W(x_j)\phi_j(y) = \sum_{j=1}^{J} W(x_j)\phi'_j(x) = \mathcal{A}'W(x).$$

These functions $\{\phi'_j := P\phi_j\}_{1\leq j\leq J}$ may be precomputed before simulations, or approximated on-line with function approximation techniques.

2.5 Other Examples

Other possible settings include the finite-horizon time, the infinite horizon stochastic shortest path, and average reward problems, briefly described now.

In a *finite-time horizon problem*, the value $V(t,x)$ is time-dependent. So let $X(t,x) = \{x_s\}_{t\leq s\leq T}$ be a trajectory starting from $x \in \mathcal{X}$ at time $t \in \{0,\dots T\}$. Write $\Psi(r,X(t,x)) := \sum_{s=t}^{T} r(x_s)$. The value function $V(t,x) = \mathbb{E}[\Psi(r,X(t,x))]$ solves Bellman's equation

$$V(t,x) = r(x) + \sum_{y\in\mathcal{X}} P(x,y)V(t+1,y), \text{ for } 0 \leq t < T$$

and $V(T,x) = r(x)$. A similar variance reduction method holds in the product space $\{0,\dots,T\} \times \mathcal{X}$. Approximate functions $\mathcal{A}W$ are defined on a grid $\{(t_j,x_j)\}_{1\leq j\leq J}$ over the product space, as a linear combination of basis functions $\{\phi_j(t,x)\}$: for any function W defined on the product space, $\mathcal{A}W(t,x) := \sum_{j=1}^{J} W(t_j,x_j)\phi_j(t,x)$. The variance reduction result of the previous section applies immediately to this case.

In *infinite horizon stochastic shortest path problems*, we usually assume that the reward function is non-negative (or non-positive if it represents a cost function) and that there exists an absorbing state (with a zero reward) that is reached, from any initial state, in finite time with probability 1. The functional is $\Psi(r,X(x)) := \sum_{t\geq 0} r(x_t)$ and the value function V solves Bellman's equation $(I-P)V = r$ with $(I-P)$ being invertible.

The case of *average reward problems* is more subtle and would deserve deeper treatment. We simply provide the idea of the possible application to this case. The functional is $\Psi(r,X(x)) := \lim_{T\to\infty} \frac{1}{T} \sum_{t=0}^{T-1} r(x_t)$. In aperiodic, recurrent, unichain Markov chains, the average expected gain ρ

$$\rho(x) := \Big(\lim_{T\to\infty} \frac{1}{T} \sum_{t=0}^{T-1} P^t r \Big)(x)$$

is independent from the start state $\rho(x) = \rho$, and satisfies $\rho = \pi r$, where π is the stationary distribution of the chain (π is considered as a row vector), i.e. $\pi P = \pi$. The relative value function $V(x) := \mathbb{E}[\Psi(r-\rho, X(x))]$ solves the equation $(I-P)V = r-\rho$. This equation has several solutions but a unique one V such that $P^{\pi}V = 0$, with P^{π} being the matrix with all rows equal to π.

In this setting, a possible extension of the variance reduction technique would process two approximations ρ_n and V_n of the average reward ρ and the relative value function V, respectively.

2.6 Numerical Experiment

We consider the *Gambler's ruin problem* described in (Kollman et al., 1999): a gambler with i dollars bets repeatedly against the house, whose initial capital is $L-i$. Each bet is one dollar and the gambler has probability p of winning. The state space is $\mathcal{X} = \{0, \dots, L\}$ and the transition matrix P is defined, for $i, j \in \mathcal{X}$, by

$$P_{ij} = \begin{cases} p, & \text{if } j-i = 1 \text{ and } 0 < i < L, \\ 1-p, & \text{if } i-j = 1 \text{ and } 0 < i < L, \\ 0, & \text{otherwise.} \end{cases}$$

Betting continues until either the gambler is ruined ($i = 0$) or he has "broken the bank" ($i = L$) (thus 0 and L are terminal states). This is an infinite-horizon time stochastic shortest path problem. We are interested in computing the probability of the gambler's eventual ruin $V(i)$ when starting from initial fortune i. We thus define the function $r(0) = 1$ and $r(i \neq 0) = 0$. The value function V solves the Bellman equation $(I-P)V = r$, and its value is

$$V(i) = \frac{\lambda^i - \lambda^L}{1-\lambda^L}, \text{ for } i \in \mathcal{X}, \tag{12}$$

with $\lambda := \frac{1-p}{p}$ when $p \neq 0.5$, and $V(i) = 1 - i/L$ for $p = 0.5$. The representative states are $X_J = \{1, 7, 13, 19\}$ (here $L = 20$). We consider two linear function approximations $\mathcal{A}_1$ and $\mathcal{A}_2$ that are projection operators (minimizing the L_2 norm at the states X_J) onto the space spanned by a set of functions $\{\psi_k : \mathcal{X} \to \mathbb{R}\}_{1 \leq k \leq K}$. $\mathcal{A}_1$ uses $K = 2$ functions $\psi_1(i) = 1, \psi_2(i) = \lambda^i, i \in \mathcal{X}$, whereas $\mathcal{A}_2$ uses $K = 4$ functions $\psi_1(i) = 1, \psi_2(i) = i, \psi_3(i) = i^2, \psi_4(i) = i^3$, $i \in \mathcal{X}$. Notice that V is representable by $\mathcal{A}_1$ (i.e. $\mathcal{A}_1 V = V$) but not by $\mathcal{A}_2$. We chose $p = 0.51$.

We ran the algorithm with $\mathcal{L} = I - P$ (which is an invertible matrix). At each iteration, we used $M = 100$ simulations per state. Figure 1 shows the L_∞ approximation error ($\max_{j \in X_J} |V(j) - V_n(j)|$) in logarithmic scale, as a function of the iteration number $1 \leq n \leq 10$. This approximation error (which is the true quantity of interest) is directly related to the variance of the estimates V_n.

For the approximation $\mathcal{A}_1$, we observe the geometric convergence to 0, as predicted in Theorem 2. It takes less than 10×100 simulations per state to reach an error of 10^{-15}. Using $\mathcal{A}_2$, the error does not decrease below some threshold $\simeq 2.10^{-5}$ due to the approximation error $V - \mathcal{A}_2 V$. This threshold is reached using about 5×100 simulations per state. For comparison, usual MC reaches an error of 10^{-4} with 10^8 simulations per state.

The variance reduction obtained when using such sequential control variates is thus considerable.

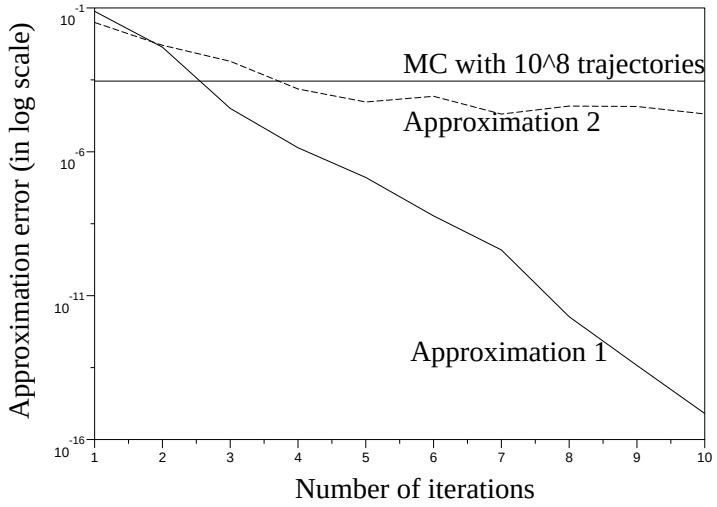

Figure 1: Approximation error for regular MC and sequential control variate algorithm using two approximations $\mathcal{A}_1$ and $\mathcal{A}_2$, as a function of the number of iterations.

3. Gradient Estimation

Here, we assume that the transition matrix P depends on some parameter α, and that we wish to estimate the sensitivity of $V(x) = \mathbb{E}[\Psi(r, X(x))]$ with respect to α, which we write $Z(x) := \partial_\alpha V(x)$.

An example of interest consists in solving approximately a Markov Decision Problem by searching for a feedback control law in a given class of parameterized stochastic policies. The optimal control problem is replaced by a parametric optimization problem, which may be solved (at least in order to find a local optimum) using gradient methods. Thus we are interested in estimating the gradient of the performance measure w.r.t. the parameter of the policy. In this example, the transition matrix P would be the transition matrix of the MDP combined with the parameterized stochastic policy.

As mentioned in the introduction, the gradient may be expressed as an expectation $Z(x) = \mathbb{E}[\Phi(r, X(x))]$ (using the so-called *likelihood ratio* or *score method* (Reiman and Weiss, 1986; Glynn, 1987; Williams, 1992; Baxter and Bartlett, 2001; Marbach and Tsitsiklis, 2003)) where $\Phi(r, X(x))$ is also a functional that depends on the trajectory $X(x)$, and that is linear in its first variable. For example, in the discounted case (3), the functional Φ is given by (5). The variance is usually high, thus variance reduction techniques are highly needed (Greensmith et al., 2005).

The gradient Z is also the solution to the linear system (4). Unfortunately, this linear expression is not of the form (2) since $\partial_\alpha \mathcal{L}$ is not invertible, which prevents us from using directly the method of the previous section.

However, the linear equation (4) provides us with another representation for Z in terms of a probabilistic representation:

$$Z(x) = \mathbb{E}[\Phi(r, X(x))] = \mathbb{E}[\Psi(-\partial_\alpha \mathcal{L} V, X(x))]. \tag{13}$$

We may extend the previous algorithm to the estimation of Z by using two representations: V_n and Z_n. The approximation V_n of V is updated from Monte-Carlo estimation of the residual $r - \mathcal{L}V_n$, and Z_n, which approximates Z, is updated from the gradient residual $-\partial_\alpha \mathcal{L}V_n - \mathcal{L}Z_n$ built from the current V_n. This approach may be related to the so-called *Actor-Critic algorithms* (Konda and Borkar, 1999; Sutton et al., 2000), which use the representation (13) with an approximation of the value function.

A geometric variance reduction is also achieved, up to a threshold that depends on the approximation errors of both of those representations.

Finally, we present a variance reduction technique that only makes use of the gradient representation Z_n (which may be useful for Partially Observable MDPs) but at the cost of a variance increase.

3.1 The Algorithm

Although the approximation operators for V and Z may be different in practice (they may use different sets of representative states and basis functions), in this section, we will use the same approximation operator $\mathcal{A}$ for simplicity.

From (13) and the equivalence property (8), we obtain the following representation for Z:

$$\begin{aligned} Z(x) &= \mathcal{A}Z_n(x) + \mathbb{E}\big[\Psi(-\partial_\alpha \mathcal{L}V - \mathcal{L}\mathcal{A}Z_n, X(x))\big] \\ &= \mathcal{A}Z_n(x) + \mathbb{E}\big[\Psi(-\partial_\alpha \mathcal{L}(V - \mathcal{A}V_n), X(x)) - \Psi(\partial_\alpha \mathcal{L}\mathcal{A}V_n + \mathcal{L}\mathcal{A}Z_n, X(x))\big] \\ &= \mathcal{A}Z_n(x) + \mathbb{E}\big[\Phi(r - \mathcal{L}\mathcal{A}V_n, X(x)) - \Psi(\partial_\alpha \mathcal{L}\mathcal{A}V_n + \mathcal{L}\mathcal{A}Z_n, X(x))\big]. \end{aligned} \tag{14}$$

from which the algorithm is deduced. We consider successive approximations $V_n \in \mathbb{R}^J$ of V and $Z_n \in \mathbb{R}^J$ of Z defined at the states $\mathcal{X}_J = (x_j)_{1 \le j \le J}$.

- We initialize $V_0(x_j) = 0$, $Z_0(x_j) = 0$.
- At stage n, we simulate by Monte Carlo M trajectories $(X^{n,m}(x_j))_{1 \le m \le M}$ and define the new approximations V_{n+1} and Z_{n+1} at the states $\mathcal{X}_J$:

$$\begin{aligned} V_{n+1}(x_j) &= \mathcal{A}V_n(x_j) + \frac{1}{M}\sum_{m=1}^{M} \Psi(r - \mathcal{L}\mathcal{A}V_n, X^{n,m}(x_j)) \\ Z_{n+1}(x_j) &= \mathcal{A}Z_n(x_j) + \frac{1}{M}\sum_{m=1}^{M} \Big[\Phi(r - \mathcal{L}\mathcal{A}V_n, X^{n,m}(x_j)) \\ &\qquad - \Psi(\partial_\alpha \mathcal{L}\mathcal{A}V_n + \mathcal{L}\mathcal{A}Z_n, X^{n,m}(x_j))\Big]. \end{aligned}$$

3.2 Properties of the Estimates V_n and Z_n

Expectation of V_n and Z_n. We have already seen that $\mathbb{E}[V_n] = V$ for all $n > 0$. Now, (14) implies that $\mathbb{E}^n[Z_{n+1}] = Z$, thus $\mathbb{E}[Z_n] = Z$ for all $n > 0$.

Variance of V_n and Z_n. We write $v_n = \sup_{1 \le j \le J} \operatorname{Var} V_n(x_j)$ and $z_n = \sup_{1 \le j \le J} \operatorname{Var} Z_n(x_j)$. The next theorem (proved in Appendix B) states the geometric variance reduction for large enough values of M.

Theorem 4 *We have*

$$\begin{aligned} v_{n+1} &\leq \rho_M v_n + \frac{2}{M}\mathcal{V}_\Psi(V - \mathcal{A}V) \\ z_{n+1} &\leq \rho_M z_n + \frac{2}{M}[c_1(V - \mathcal{A}V, Z - \mathcal{A}Z) + c_2 v_n] \end{aligned}$$

with $\rho_M = \frac{2}{M}\left(\sum_{j=1}^J \sqrt{\mathcal{V}_\Psi(\phi_j)}\right)^2$, *and the coefficients*

$$\begin{aligned} c_1(f,g) &= \left(\sqrt{\mathcal{V}_\Phi(f)} + \sqrt{\mathcal{V}_\Psi(\mathcal{L}^{-1}\partial_\alpha \mathcal{L} f)} + \sqrt{\mathcal{V}_\Psi(g)}\right)^2 \\ c_2 &= \Big[\sum_{j=1}^J \sqrt{\mathcal{V}_\Phi(\phi_j)} + \sqrt{\mathcal{V}_\Psi(\mathcal{L}^{-1}\partial_\alpha \mathcal{L}\phi_j)}\Big]^2, \end{aligned}$$

using the notations $\mathcal{V}_\Psi(f) := \sup_{1\leq j\leq J} Var\ \Psi(\mathcal{L}f, X(x_j))$ *and* $\mathcal{V}_\Phi(f) := \sup_{1\leq j\leq J} Var\ \Phi(\mathcal{L}f, X(x_j))$. *Thus, for large enough values of M, (i.e. whenever* $\rho_M < 1$*), the convergence of* $(v_n)_n$ *and* $(z_n)_n$ *is geometric at rate* ρ_M*, up to the thresholds*

$$\begin{aligned} \limsup_{n\to\infty} v_n &\leq \frac{1}{1-\rho_M}\frac{2}{M}\mathcal{V}_\Psi(V - \mathcal{A}V) \\ \limsup_{n\to\infty} z_n &\leq \frac{1}{1-\rho_M}\frac{2}{M}\Big[c_1(V - \mathcal{A}V, Z - \mathcal{A}Z) + c_2\frac{1}{1-\rho_M}\frac{2}{M}\mathcal{V}_\Psi(V - \mathcal{A}V)\Big]. \end{aligned}$$

Here also, if V and Z are representable by $\mathcal{A}$, then the variance converges geometrically to 0.

3.3 Numerical Experiment

Again we consider the *Gambler's ruin problem* described previously. The transition matrix is parameterized by $\alpha = p$, the probability of winning. The gradient $Z(i) = \partial_\alpha V(i)$ may be derived from (12):

$$Z(i) = \frac{L(1-\lambda^i)\lambda^{L-1} - i(1-\lambda^L)\lambda^{i-1}}{(1-\lambda^L)^2\alpha^2} \text{ for } i \in \mathcal{X},$$

(for $\alpha \neq 0.5$), and $Z(i) = 0$ for $\alpha = 0.5$. Again we use the representative states $X_J = \{1, 7, 13, 19\}$. Here, we consider two possible approximators $\mathcal{A}_1$ and $\mathcal{A}_2$ for the value function representations V_n (as defined previously), and two approximators $\mathcal{A}_2$ and $\mathcal{A}_3$ for the gradient representations Z_n, where $\mathcal{A}_3$ is the projection that uses $K = 3$ functions $\psi_1(i) = 1, \psi_2(i) = \lambda^i, \psi_3(i) = i\lambda^i, i \in \mathcal{X}$. Notice that Z is representable by $\mathcal{A}_3$ but not by $\mathcal{A}_2$. We choose $p = 0.51$ and $M = 1000$.

Figure 2 shows the L_∞ approximation error of Z ($\max_{j\in X_J} |Z(j) - Z_n(j)|$) in logarithmic scale, for the different possible approximations of V and Z.

When both V and Z may be perfectly approximated (i.e. $\mathcal{A}_1$ for V and $\mathcal{A}_3$ for Z) we observe the geometric convergence to 0, as predicted in Theorem 4. The error is around 10^{-14} using a total of 10^4 simulations. When either the value function or the gradient is not representable in the approximation spaces, the error does not decrease below some threshold ($\simeq 3.10^{-3}$ when Z is not representable) reached in 2.10^3 simulations. The threshold is lower ($\simeq 2.10^{-4}$) when Z is representable. For comparison, usual MC reaches an error (for Z) of 3.10^{-3} with 10^8 simulations per state.

The variance reduction of this sequential method compared to regular MC is thus also considerable.

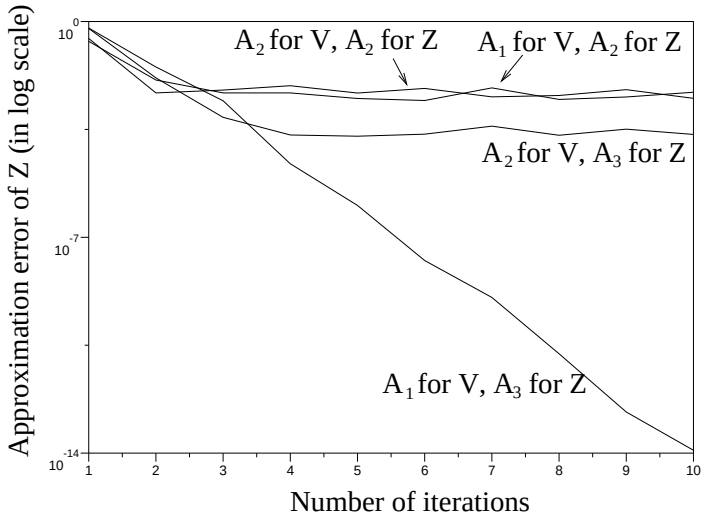

Figure 2: Approximation error of the gradient $Z = \partial_\alpha V$ using approximators $\mathcal{A}_1$ and $\mathcal{A}_2$ for the value function, and $\mathcal{A}_2$ and $\mathcal{A}_3$ for the gradient.

3.4 Variance Reduction with Only Z Representation

It would be desirable to design a similar variance reduction method using the gradient approximation only. However, as seen previously, the linear system (4) does not enable to recover r from the gradient (since $\partial_\alpha \mathcal{L}$ is not invertible), which prevents us from directly using the method of Section 2.

Nevertheless, from (13), we have the representation for Z:

$$Z(x) \quad = \quad \mathcal{A}Z_n(x) + \mathbb{E}\big[\Phi(r,X(x)) - \Psi(\mathcal{L}\mathcal{A}Z_n, X(x))\big],$$

from which we deduce the following algorithm: at stage n, simulate M trajectories $X^{n,m}$ per state (x_j) and update the approximation Z_n according to

$$Z_{n+1}(x_j) \quad = \quad \mathcal{A}Z_n(x_j) + \frac{1}{M}\sum_{m=1}^{M}\Big[\Phi(r,X^{n,m}(x_j)) - \Psi(\mathcal{L}\mathcal{A}Z_n, X^{n,m}(x_j))\Big].$$

Unfortunately, we may not expect this algorithm to exhibit a variance reduction to 0 in the case of perfect approximation of the gradient (i.e. $\mathcal{A}Z = Z$). Indeed, there is an incompressible variance term that comes from the estimation of $\Phi(r,X(x))$ instead of $\Psi(\mathcal{L}Z,X(x)) = \Psi(-\partial_\alpha \mathcal{L}\,\mathcal{L}^{-1}r, X(x))$.

To illustrate, in the infinite-horizon, discounted case (5), this incompressible variance term appears in the estimation of

$$\Phi(r,X(x)) - \Psi(\mathcal{L}Z,X(x)) \quad = \quad \sum_{t\geq 0}\gamma^t\Big[\frac{\partial_\alpha P(x_t,x_{t+1})}{P(x_t,x_{t+1})}\sum_{s\geq 0}\gamma^{s+1}r(x_{s+t+1}) - (I-\gamma P)Z(x_t)\Big].$$

However this variance (which can be related to the variance of the value function $V(x_{t+1})$ estimation by the sum of future rewards $\sum_{s\geq 0}\gamma^s r(x_{s+t+1})$ and a bound on the likelihood ratios $\frac{\partial_\alpha P(x_t,x_{t+1})}{P(x_t,x_{t+1})}$) is much lower (especially when γ is close to 1) than the initial variance of the direct estimation of $\mathbb{E}[\Phi(r,X(x))]$.

Thus, this algorithm would provide a geometric variance reduction, up to a threshold that depends on $\mathcal{V}_\Psi(Z-\mathcal{A}Z)$ plus this incompressible variance term (the proof is a simple extension of that of Theorem 2 taking into account the additional variance term). This algorithm may be interesting in Partially Observable MDPs, and provide an alternative technique compared to other variance reduction techniques developed in this setting (Greensmith et al., 2005).

4. Conclusion

We described a sequential control variates method for estimating the expectation of functionals in Markov chains, using linear approximation (in the values). We illustrate the method on value function and gradient estimates. We proved geometric variance reduction up to a threshold that depends on the approximation error of the functions of interest.

There are several possible directions for future research, among which:

- Estimate the number of sample trajectories M per state that enables the method to exhibit a geometric variance reduction (i.e. whenever $\rho_M < 1$).
- For a total budget of N trajectories per state, define what is the best trade-off between the number of iterations n and the number of trajectories M per iteration (such that $N = nM$).
- Define a stopping criterion (i.e. whenever there is no more variance decrease) from which we should continue (if needed) with a regular Monte Carlo method.
- Consider the case where the initial states are drawn according to some distribution over $\mathcal{X}$ instead of using the set of representative states $\mathcal{X}_J$.
- Consider non-linear function approximation.
- Extend this work to a model-free, on-line learning framework.

Appendix A. Proof of Theorem 2

From the decomposition

$$V - \mathcal{A}V_n = V - \mathcal{A}V + \sum_{i=1}^{J}(V-V_n)(x_i)\phi_i, \tag{15}$$

we have

$$\begin{aligned} V_{n+1}(x_j) &= \mathcal{A}V_n(x_j) + \frac{1}{M}\sum_{m=1}^{M}\Big[\Psi(\mathcal{L}(V-\mathcal{A}V), X^{n,m}(x_j)) \\ &\quad + \sum_{i=1}^{J}(V-V_n)(x_i)\Psi(\mathcal{L}\phi_i, X^{n,m}(x_j))\Big]. \end{aligned}$$

Thus

$$\begin{aligned}\mathrm{Var}^n V_{n+1}(x_j) &= \frac{1}{M}\mathrm{Var}^n\big[\Psi(\mathcal{L}(V-\mathcal{A}V),X(x_j)) \\ &\quad +\sum_{i=1}^{J}(V-V_n)(x_i)\Psi(\mathcal{L}\phi_i,X(x_j))\big].\end{aligned}$$

We use the general bound

$$\begin{aligned}\mathrm{Var}\,\big[\sum_i \alpha_i Y_i\big] &= \sum_{i_1,i_2}\alpha_{i_1}\alpha_{i_2}\mathrm{Cov}(Y_{i_1},Y_{i_2}) \\ &\le \sum_{i_1,i_2}|\alpha_{i_1}||\alpha_{i_2}|\sqrt{\mathrm{Var}\,[Y_{i_1}]}\sqrt{\mathrm{Var}\,[Y_{i_2}]} \le \big[\sum_i|\alpha_i|\sqrt{\mathrm{Var}\,[Y_i]}\big]^2, \qquad (16)\end{aligned}$$

for any real numbers $(\alpha_i)_i$ and square integrable real random variables $(Y_i)_i$, to deduce that

$$\mathrm{Var}^n V_{n+1}(x_j) \le \frac{1}{M}\Big[\sqrt{\mathcal{V}_\Psi(V-\mathcal{A}V)}+\sum_{i=1}^{J}|V-V_n|(x_i)\sqrt{\mathcal{V}_\Psi(\phi_i)}\Big]^2, \qquad (17)$$

with $\mathcal{V}_\Psi(f) := \sup_{1\le j\le J}\mathrm{Var}\,\Psi(\mathcal{L}f,X(x_j))$. Now, we use the variance decomposition

$$\begin{aligned}\mathrm{Var}\,V_{n+1}(x_j) &= \mathrm{Var}\,[\mathbb{E}^n[V_{n+1}(x_j)]]+\mathbb{E}[\mathrm{Var}^n[V_{n+1}(x_j)]] \\ &= \mathbb{E}[\mathrm{Var}^n[V_{n+1}(x_j)]],\end{aligned}$$

and the general bound (deduced similarly to (16))

$$\mathbb{E}\big[(\alpha_0+\sum_{i=1}^{J}\alpha_i Y_i)^2\big] \le 2\alpha_0^2+2\Big(\sum_{i=1}^{J}|\alpha_i|\sqrt{\mathbb{E}[Y_i^2]}\Big)^2, \qquad (18)$$

to deduce from (17) that

$$\nu_{n+1} \le \frac{2}{M}\big[\mathcal{V}_\Psi(V-\mathcal{A}V)+\big(\sum_{i=1}^{J}\sqrt{\mathcal{V}_\Psi(\phi_i)}\big)^2\nu_n\big],$$

which gives (10). Now, if M is such that $\rho_M := \frac{2}{M}\big(\sum_{i=1}^{J}\sqrt{\mathcal{V}_\Psi(\phi_i)}\big)^2 < 1$, then taking the upper limit finishes the proof of Theorem 2.

Appendix B. Proof of Theorem 4

Using (4) and (6), we have the decomposition

$$\begin{aligned}-\partial_\alpha\mathcal{L}\mathcal{A}V_n-\mathcal{L}\mathcal{A}Z_n &= -\partial_\alpha\mathcal{L}\mathcal{A}(V_n-V)-\partial_\alpha\mathcal{L}(\mathcal{A}V-V) \\ &\quad +\mathcal{L}(Z-\mathcal{A}Z)+\mathcal{L}\mathcal{A}(Z-Z_n) \\ &= \sum_{i=1}^{J}(V-V_n)(x_i)\partial_\alpha\mathcal{L}\phi_i-\partial_\alpha\mathcal{L}(\mathcal{A}V-V) \\ &\quad +\mathcal{L}(Z-\mathcal{A}Z)+\sum_{i=1}^{J}(Z-Z_n)(x_i)\mathcal{L}\phi_i.\end{aligned}$$

Now, using (15), the variance may be written

$$\begin{aligned}\mathrm{Var}^n Z_{n+1}(x_j) &= \frac{1}{M}\mathrm{Var}^n\Big[\Phi(\mathcal{L}(V-\mathcal{A}V),X(x_j))\\ &+\sum_{i=1}^{J}(V-V_n)(x_i)\Phi(\mathcal{L}\phi_i,X(x_j))-\Psi(\partial_\alpha\mathcal{L}(\mathcal{A}V-V),X(x_j))\\ &+\sum_{i=1}^{J}(V-V_n)(x_i)\Psi(\partial_\alpha\mathcal{L}\phi_i,X(x_j))+\Psi(\mathcal{L}(Z-\mathcal{A}Z),X(x_j))\\ &+\sum_{i=1}^{J}(Z-Z_n)(x_i)\Psi(\mathcal{L}\phi_i,X(x_j))\Big].\end{aligned}$$

We use (16) to deduce the bound

$$\begin{aligned}\mathrm{Var}^n Z_{n+1}(x_j) &\le \frac{1}{M}\Big[\sqrt{\mathcal{V}_\Phi(V-\mathcal{A}V)}+\sqrt{\mathcal{V}_\Psi(\mathcal{L}^{-1}\partial_\alpha\mathcal{L}(\mathcal{A}V-V))}\\ &+\sum_{i=1}^{J}|V-V_n|(x_i)\big(\sqrt{\mathcal{V}_\Phi(\phi_i)}+\sqrt{\mathcal{V}_\Psi(\mathcal{L}^{-1}\partial_\alpha\mathcal{L}\phi_i)}\big)\\ &+\sqrt{\mathcal{V}_\Psi(Z-\mathcal{A}Z)}+\sum_{i=1}^{J}|Z-Z_n|(x_i)\sqrt{\mathcal{V}_\Psi(\phi_i)}\Big]^2,\end{aligned}$$

Now, we use (18) to deduce that

$$\begin{aligned}z_{n+1} &\le \frac{2}{M}\Big\{\big(\sqrt{\mathcal{V}_\Phi(V-\mathcal{A}V)}+\sqrt{\mathcal{V}_\Psi(\mathcal{L}^{-1}\partial_\alpha\mathcal{L}(\mathcal{A}V-V))}\\ &+\Big[\sum_{i=1}^{J}\sqrt{\mathcal{V}_\Phi(\phi_i)}+\sqrt{\mathcal{V}_\Psi(\mathcal{L}^{-1}\partial_\alpha\mathcal{L}\phi_i)}\Big]^2 v_n\\ &+\sqrt{\mathcal{V}_\Psi(Z-\mathcal{A}Z)}+\Big[\sum_{i=1}^{J}\sqrt{\mathcal{V}_\Psi(\phi_i)}\Big]^2 z_n\Big\},\end{aligned}$$

and Theorem 4 follows.

References

C. G. Atkeson, S. A. Schaal, and Andrew W. Moore. Locally weighted learning. *AI Review*, 11, 1997.

J. Baxter and P. L. Bartlett. Infinite-horizon gradient-based policy search. *Journal of Artificial Intelligence Research*, 15:319–350, 2001.

P. W. Glynn. Likelihood ratio gradient estimation: an overview. In A. Thesen, H. Grant, and W. D. Kelton, editors, *Proceedings of the 1987 Winter Simulation Conference*, pages 366–375, 1987.

E. Gobet and S. Maire. Sequential control variates for functionals of Markov processes. *SIAM Journal on Numerical Analysis*, 43(3):1256–1275, 2005.

E. Greensmith, P. L. Bartlett, and J. Baxter. Variance reduction techniques for gradient estimates in reinforcement learning. *Journal of Machine Learning Research*, 5:1471–1530, 2005.

J. H. Halton. A retrospective and prospective survey of the Monte-Carlo method. *SIAM Review*, 12 (1):1–63, 1970.

J. H. Halton. Sequential Monte-Carlo techniques for the solution of linear systems. *Journal of Scientific Computing*, 9:213–257, 1994.

J. M. Hammersley and D. C. Handscomb. *Monte-Carlo Methods*. Chapman and Hall, 1964.

T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning*. Springer Series in Statistics, 2001.

C. Kollman, K. Baggerly, D. Cox, and R. Picard. Adaptive importance sampling on discrete Markov chains. *The Annals of Applied Probability*, 9(2):391–412, 1999.

V. R. Konda and V. S. Borkar. Actor-critic-type learning algorithms for Markov decision processes. *SIAM Journal of Control and Optimization*, 38:1:94–123, 1999.

S. Maire. An iterative computation of approximations on Korobov-like spaces. *J. Comput. Appl. Math.*, 54(6):261–281, 2003.

P. Marbach and J. N. Tsitsiklis. Approximate gradient methods in policy-space optimization of Markov reward processes. *Journal of Discrete Event Dynamical Systems*, 13:111–148, 2003.

M. I. Reiman and A. Weiss. Sensitivity analysis via likelihood ratios. In J. Wilson, J. Henriksen, and S. Roberts, editors, *Proceedings of the 1986 Winter Simulation Conference*, pages 285–289, 1986.

R. S. Sutton, D. McAllester, S. Singh, and Y. Mansour. Policy gradient methods for reinforcement learning with function approximation. *Neural Information Processing Systems. MIT Press*, pages 1057–1063, 2000.

V. Vapnik. *Statistical Learning Theory*. John Wiley & Sons, New York, 1998.

V. Vapnik, S. E. Golowich, and A. Smola. Support vector method for function approximation, regression estimation and signal processing. *In Advances in Neural Information Processing Systems*, pages 281–287, 1997.

R. J. Williams. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine Learning*, 8:229–256, 1992.